METHOD TO PROTECT DATA ON A DISK DRIVE FROM UNCORRECTABLE MEDIA ERRORS

BACKGROUND OF THE INVENTION

Field of the Invention

[01] The present invention relates to storage systems. More particularly, the present invention relates to a system, a method and a storage format that provides protection against uncorrectable media errors.

Description of the Related Art

[02] Figure 1 shows an exemplary high-RPM Hard Disk Drive (HDD) 100 having a two-stage servo system for positioning a magnetic read/write head (or recording slider) 101 over a selected track on a magnetic disk 102. The two-stage servo system includes a voice-coil motor (VCM) 103 for coarse position a read/write head suspension 104 and a microactuator, or micropositioner, for fine positioning read/write head 101 over the selected track in a well-known manner. Binary data is stored on magnetic disk 102 by selectively orienting magnetization in user data fields in the magnetic media of disk 102.

[03] The two primary sources of data loss from HDDs, such as the exemplary HDD shown in Figure 1, are disk drive failure and uncorrectable media error. Data loss has been conventionally prevented by configuring storage systems having an array of multiple HDDs in a RAID configuration in which data is striped across multiple HDDs. Redundancy is built into the striping so that should any HDD fail, the data belonging to the failed HDD can be reconstructed from the remaining drives of the storage system.

[04] HDD storage capabilities have been increasing at a rate of between 60 and 100 percent per year. The probability of uncorrectable read errors, however, has been relatively constant at

about 1 uncorrectable read error in 10^{14} bits. Accordingly, as HDD storage capabilities have increased, the probability of data loss due to an uncorrectable media error has become a significant factor.

[05] Multiple HDD storage systems configured as RAID level 5 systems are commonly deployed in the industry and can tolerate loss of a single disk HDD. While a failed HDD is being rebuilt, however, a second HDD failure or an uncorrectable media error on any of the remaining HDDs will result in data loss. Data loss caused by a second HDD failure is referred to as an "array loss," while data loss caused by an uncorrectable media error is referred to as a "strip kill." It is estimated that there will be 1.48 array losses and 2570 strip kills in a one-year period for an installed base of one million 300 GB HDDs that are configured in 8-drive RAID 5 array systems with each HDD having an MTBF of 500,000 hours. It should be noted that over 90% of media errors affect single sectors. About 5% of media errors affect two to four sectors. Very few media errors affect multiple (seven or more) sectors.

[06] Techniques have been proposed for reducing the probability of data loss. In particular, RAID-type protection techniques have been developed for protecting against drive failure by increasing the redundancy of the array using levels (such as RAID 51, RAID 6, RAID (3+3) and so on). When a RAID level is chosen for a storage system, factors that are considered include storage efficiency, reliability and performance. Optimizing any one of these three factors causes at least one of the other factors to become less than optimal.

[07] Table 1 is a comparison of the different conventional RAID techniques.

Table 1

	RAID 5	RAID 51	RAID (3+3)	RAID 6	RAID N+3
Drives/array	8	16	6	16	16
Storage Efficiency	87.5%	43.75%	50%	87.5%	81.25%
Annual Strip Kill events	2570	6.17×10^{-7}	8.55×10^{-7}	1.61	7.53×10^{-4}
Annual Array Loss events	1.48	5.01×10^{-8}	3.56×10^{-10}	2.41×10^{-3}	1.51×10^{-6}
Performance (IOs/writes)	4	6	6	6	8

[08] The parameters on which the reliability calculations in Table 1 are based are an installed base of one million 300 GB disk drives each having an MTBF of 500,000 hours and a hard error rate of 1 error in 10^{14} bits.

[09] As can be seen from Table 1, a RAID 6 system configuration provides an adequate protection against array loss events exhibiting only 2.41×10^{-3} array loss events per year. The number of strip kills (i.e., 1.61 strip kill events per year) is too many to meet the requirements of high-end storage systems. Adding another level of protection comes at a price, such as reduced storage efficiency (i.e., a RAID (3+3) or a RAID 51 system configuration) or reduced performance (i.e., a RAID N+3 system configuration).

[10] While RAID-type protection techniques have been developed for protecting against drive failure, RAID techniques do not protect well against uncorrectable media error and result in

coarse granularity and sub-optimal tradeoffs. Consequently, what is needed is a technique that provides protection against uncorrectable media errors.

BRIEF SUMMARY OF THE INVENTION

[11] The present invention provides a technique that provides protection against uncorrectable media errors.

[12] The advantages of the present invention are provided by a method and a system for protecting data stored in a RAID-configured storage system from uncorrectable media errors. The RAID-configured storage system includes a plurality of storage units, such as HDDs, optical drives and/or random access memory. According to the invention, c redundancy information sectors are associated with n data information sectors, such that the c redundancy information sectors are based on the n data information sectors. The n data information sectors are written with c redundancy information sectors in a data segment on the same storage unit. The RAID-configured storage system can be configured, for example, as a RAID 6 storage system, a RAID 5 storage system, a RAID 51 storage system, a RAID 3 + 3 storage system or a RAID N + 3 storage system. The c redundancy sectors effectively protect the n information sectors against up to c uncorrectable media errors in the n sectors. The redundancy information can be based on a Reed-Solomon code, an XOR-based code, or one-dimensional parity. The n data information sectors and the c redundancy information sectors can be arranged to be consecutive or intermingled.

[13] The present invention also provides a storage medium having a recording format therein having c redundancy information sectors that are associated with n data information sectors to form a segment. The c redundancy information sectors are based on the n data information sectors, and the segment is stored on a single storage unit in an array of storage units in a RAID-configured storage system.

BRIEF DESCRIPTION OF THE DRAWINGS

[14] The present invention is illustrated by way of example and not by limitation in the accompanying figures in which like reference numerals indicate similar elements and in which:

[15] Figure 1 shows an exemplary high-RPM Hard Disk Drive;

[16] Figure 2 shows an exemplary array arranged in a RAID 6 system configuration; and

[17] Figure 3 depicts an exemplary format arrangement of n data sectors and c code sectors according to the present invention.

DETAILED DESCRIPTION OF THE INVENTION

[18] The present invention provides protection against uncorrectable media errors by writing data and redundancy information on the same disk drive using a technique referred to as SPIDRE (Sector Protection through Intra Disk REdundancy).

[19] Figure 2 shows an exemplary array 200 of six storage units, such as HDDs, arranged in a RAID 6 system configuration. For the exemplary RAID 6 system configuration shown in Figure 2, parity is calculated based on data blocks arranged horizontally across storage units 0-6, similar to a RAID 5 system configuration, with a second set of parity that is also calculated based on data blocks arranged horizontally across the storage units. The first horizontal parity block is calculated, for example, as the XOR of the data blocks. The second horizontal parity block can, for example, be based on a Reed-Solomon code. Specifically, parity block P0 is based on data blocks D0-D3. Parity block P1 is also based on data blocks D0-D3. Parity blocs P2 and P3 are based on data blocks D4-D7. Parity blocks P4 and P5 are based on data blocks D8-D11. Parity blocks P6 and P7 are based on data blocks D12-D15. Parity blocks P8 and P9 are based on data blocks D16-D19. Parity blocks P10-P11 are based on data blocks D20-D24. An array controller 210 is commonly connected to all storage units

in array 200. Array controller 210 communicates with other controllers and host systems (not shown) over interface 211. Array controller 210 may be designed as a hardware and/or a software controller.

[20] According to the invention, a segment of n data sectors is associated with a set of c code (correction code or SPIDRE code) sectors. Figure 3 depicts an exemplary format arrangement of n data sectors and c code sectors according to the present invention. The n data sectors and the c redundancy sectors are written together on a single storage unit, such as storage unit A0 in Figure 2. The c sectors protect against uncorrectable media errors up to c sectors within the given data segment. There is no requirement that the n data sectors and the c redundancy sectors be kept separate. By keeping them separate, however, normal read operations are simple and fast.

[21] The present invention provides the advantage of flexible tradeoffs in storage efficiency, performance and reliability, which can be optimized by properly selecting values for n and c. For example, when c is selected to be 10% of n, the resulting storage efficiency is about 91%. Additionally, the performance impact when c code sectors are written is minimized because no seek operation is required. The data and SPIDRE code are written together. Moreover, while write operations for the present invention require a read-modify-write operation, these are, however, performed for a RAID-configured system that provides protection against drive failures "above" the protection provided by the present invention. Thus, the overhead associated with the SPIDRE codes of the present invention involves only the writing of the n + c sectors, instead of writes of the requested sector, which occur for a worst case single-sector write. For a 10,000 rpm HDD having 350 Kbytes per track, reading or writing 64 Kbytes takes about 1 ms. Data reliability is significantly enhanced, particularly when the technique of the present invention is used in conjunction with a technique such as RAID 6.

[22] For example, a code length of 8 sectors over a segment size of 128 sectors (64 Kbytes) will have a storage efficiency of 94% (82.25% when used in a RAID 6-configured system) and have a performance overhead of about 12% when used with 10,000 rpm drives. The number of annual strip kill events will drop from 1.61 to 2.21 x 10^{-3}. By selecting proper values of data segment size and code length, the three main parameters affecting RAID system design selection – storage efficiency, reliability and performance – can be optimized at a granularity not available with conventional RAID system configurations alone, which are primarily intended to deal with drive failure.

[23] Table 2 shows a comparison of a RAID 6 system configuration using SPIDRE codes according to the present invention with the other RAID configurations set forth in Table 1

Table 2

	RAID 5	RAID 51	RAID (3+3)	RAID 6	RAID N+3	RAID 6 w/SPIDRE
Drives/array	8	16	6	16	16	16
Storage Efficiency	87.5%	43.75%	50%	87.5%	81.25%	82.25%
Annual Strip Kill events	257	6.17 x 10^{-8}	3.56 x 10^{-8}	0.16	7.56 x 10^{-5}	2.21 x 10^{-3}
Annual Array Loss events	1.48	5.01 x 10^{-8}	3.56 x 10^{-10}	2.41 x 10^{-3}	1.51 x 10^{-6}	2.41 x 10^{-3}
Performance (IOs/write)	4	6	6	6	8	6.72

[24] While SPIDRE codes are illustrated in connection with a RAID 6 system configuration, SPIDRE codes can be used with a system configured for any RAID level, such as RAID 5 and RAID 0. Regardless of the RAID system configuration that the present invention is used with, the combined reliability of the present invention in conjunction with the underlying RAID-configured system is significantly higher than for the RAID-configured system alone.

[25] Several different types of erasure codes can to be used for SPIDRE codes. For example, a Reed-Solomon erasure code, which is a general code, can be selected as a SPIDRE code for protecting any combination of n and c. Alternatively, an XOR-based code can be used for large values of n. Variations of one-dimensional parity can be used as yet another alternative type of erasure code. For instance, one sector from each group of, for example, 8 sectors, is XORed with corresponding sectors from other groups. This alternative approach is simple and protects against media errors of consecutive sectors up to the group size.

[26] While the present invention has been described in terms of storage arrays formed from storage units, such as HDDs, the present invention is applicable to storage systems formed from arrays of other memory devices, such as Random Access Memory (RAM) storage devices (both volatile and non-volatile), optical storage devices, and tape storage devices. Additionally, it is suitable to virtualized storage systems, such as arrays built out of network-attached storage.

[27] Although the foregoing invention has been described in some detail for purposes of clarity of understanding, it will be apparent that certain changes and modifications may be practiced that are within the scope of the appended claims. Accordingly, the present embodiments are to be considered as illustrative and not restrictive, and the invention is not to be limited to the details given herein, but may be modified within the scope and equivalents of the appended claims.